Elderton
Primer of Statistics

PRIMER OF STATISTICS

BY

W. PALIN ELDERTON
FELLOW OF THE INSTITUTE OF ACTUARIES
AND
ETHEL M. ELDERTON
GALTON RESEARCH SCHOLAR IN NATIONAL EUGENICS

LONDON
ADAM AND CHARLES BLACK
1910

NOTE

WE wish to express our thanks to Sir Francis Galton for the original suggestion of this primer and for much kind help and interest throughout its preparation; to Professor Karl Pearson for suggestions and advice on many points; and to Mr. C. A. Sutton for reading the manuscript from the lay point of view.

W. P. E.
E. M. E.

First Edition published November, 1909.
Second Edition, May, 1910.

PREFACE

BY SIR FRANCIS GALTON, D.C.L., F.R.S.

IN my 'Herbert Spencer' lecture of 1907 before the University of Oxford, I expressed a belief that the elementary ideas on which the modern system of statistics depend, that the quality of the results to which they lead, and that the meaning of the uncouth words used in their description, admitted of much simpler explanation than usual. I sketched out a possible course of lectures to be accompanied with certain simple sortings, with object lessons and with diagrams. Finally, I expressed the hope that some competent teacher would elaborate a course of instruction on these lines. I entertain a strong belief that such a course would be of great service to those who are interested in statistics, but who, from want of mathematical aptitudes and special study, are unable to comprehend the results arrived at, even as regards their own subjects. It is, for example, a great hindrance to have no knowledge of what is meant by 'correlation.'

I learnt with much pleasure that two very competent persons were disposed to undertake the task —namely, Mr. W. Palin Elderton, well known as a highly instructed actuary, and his sister, Miss Ethel M. Elderton, who holds the post of Research Scholar in the Eugenics Laboratory of the University of London (now located in University College), and who is a thoroughly experienced worker in the modern methods.

This primer is the result. It goes forth on its important errand of familiarizing educated persons with the most recent developments of the new school of statistics, and, I beg to be allowed to add, with my heartiest good wishes for its success.

September, 1909.

CONTENTS

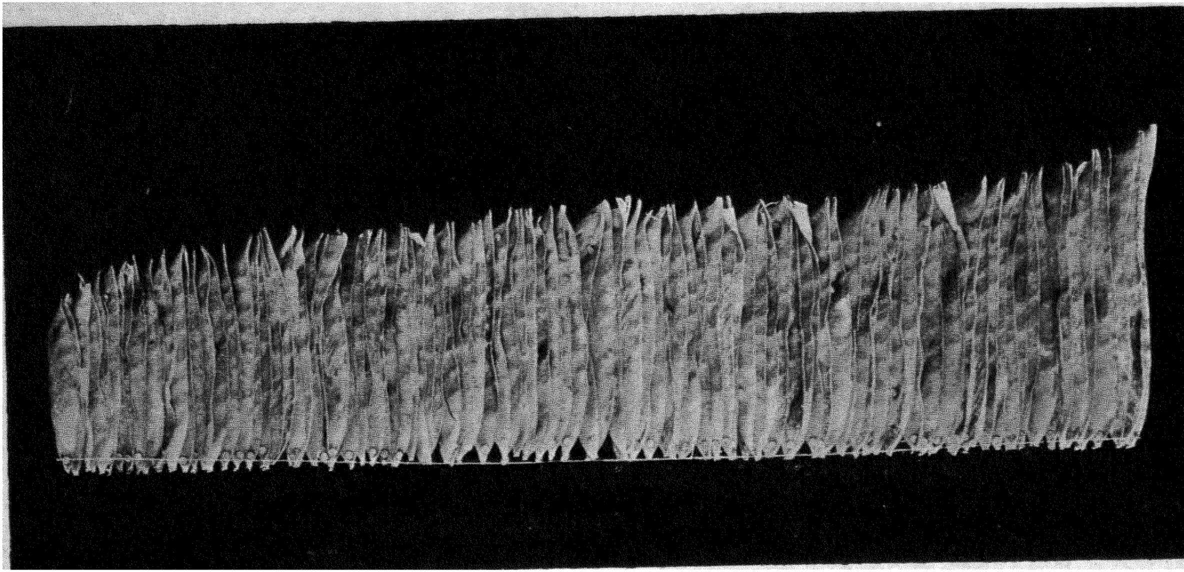

PLATE SHOWING PEAPODS ARRANGED IN ORDER OF LENGTH.

(NOTE.—*The pods are arranged on a straight line; this is unfortunately obscured, owing to the thread or cotton shown in the plate having shifted.*)

To face p. 1.

PRIMER OF STATISTICS

CHAPTER I

VARIATES AND MEDIANS

If you will go into the garden and pick a number of leaves from a tree, you will find that they are not all of the same size; some leaves, even though they are fully grown, will be much smaller than others, and if you were to measure their lengths, you would find a considerable difference between the longest and shortest specimens in your collection. You would probably obtain similar results in any other class of objects you collected, and if you were asked what was the size of the leaf of a particular tree, or the size of a certain kind of nut or shell, you would have to reply that, as you found all sorts of sizes, you could not give an exact answer. You could explain that these things always varied because of the many causes that affect size; but if you thought the matter over, you would not be long in seeing that such an explanation did not supply

a full answer to the question, but merely accentuated your inability to give one. Let us see if we can find some way of working out an answer, and of expressing the results of your measurements in an intelligible form.

It does not matter very much what class of objects is chosen, but it is well to start with something that can be easily handled and measured, and with this in view we will first take some actual measurements of shells. If you empty the shells from a bag, they will not fall in any particular

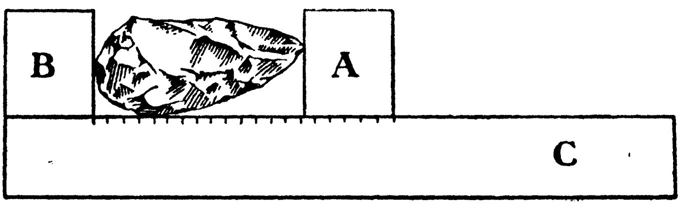

FIG. 1.

order, but admit of being arranged, and this can be done in a great many ways, as in the order of their lengths or breadths or their weights. Let us take the lengths of the shells as a first example, and begin by measuring all the specimens. A convenient way to do this is with an instrument* such as that pictured above (Fig. 1), in which C is

* The reader can easily make such an instrument for rough measurements. He wants an even piece of wood for C, on which he must paste a piece of paper ruled in tenths of an inch (say), and B and A can be made with old match-boxes or pieces of wood.

a long scale, B a fixed block, and A a movable one. The object is placed on the scale touching B, and A is shifted along until it touches the other end of the object. The length is then read off.

A moderate number (59) of shells were each measured in this way. They were then placed side

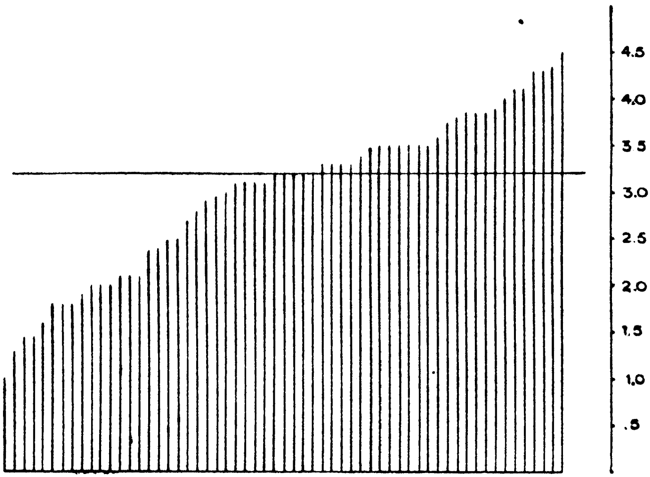

FIG. 2.—SHOWING LENGTHS OF 59 SHELLS.

by side, in the order of their lengths, at equal distances apart, or, as it is called, were 'arrayed' in order of their lengths. If we draw at equal distances along a horizontal base vertical lines having the same lengths as those of the shells, we shall get an array of upright lines like those in Fig. 2. A line joining the tops of these uprights runs evenly enough to suggest that, when shells are

properly arranged, there is probably some way of showing that their lengths do not vary quite irre-

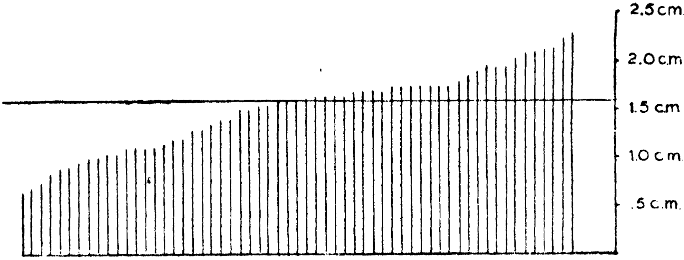

FIG. 3 —SHOWING BREADTHS OF 59 SHELLS.

gularly, but that their variations can be expressed as forming a regular series.

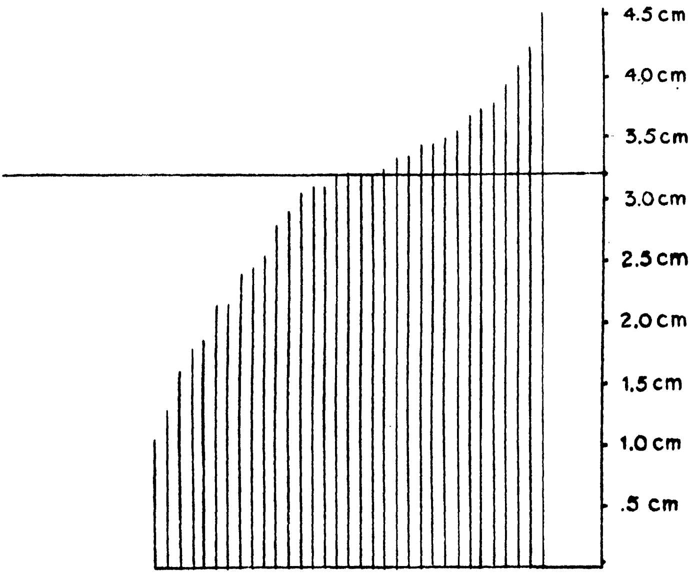

FIG. 4.—SHOWING LENGTHS OF 33 SHELLS.

This is the first step towards finding that the variations of the lengths of shells are governed by law; but it will immediately occur to you that, though the population of shells in this particular collection varies regularly, that of another sample of the same sort of shells might give a very different appearance or run unevenly. In order to investigate this, it is necessary to make some further sets of measurements, and see if a similar result occurs.

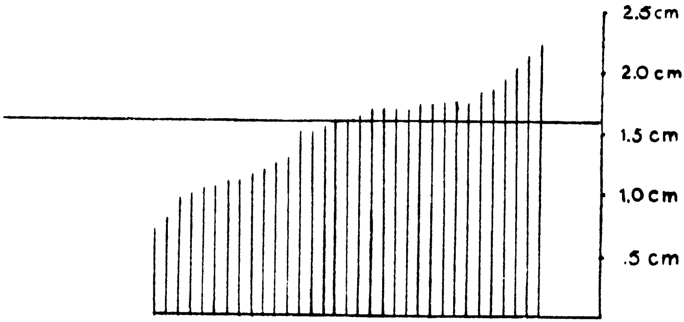

FIG. 5.—SHOWING BREADTHS OF 33 SHELLS.

We will take another collection containing fewer shells (say 33), and arrange them in order of their greatest lengths, just as we did in the previous lot. They will form arrays like those in Fig. 4.

We will also arrange the two series of 59 and 33 shells in the order of their breadths (Figs. 3 and 5).

Now examine the diagrams, and notice any points of similarity you can. (1) All the diagrams are

flat near the middle shell, or, in other words, a great many shells, or 'variates,' in each collection are of about the same size as that of the middle shell, or variate. (2) This middle variate is of nearly the same size in each case, as you can see by looking at the heights at which the horizontal lines are drawn to distinguish them. Statisticians call this middle term the 'median,' so that we can say that the median length of shells is 3·23 cm. in the first series, 3·20 cm. in the second. You will notice that these values differ but little. (3) It is also fairly clear that the diagrams have the same characteristic shape; they all rise steeply at either end, and are flat in the middle, which means that big differences ('deviations') from the middle term, in the lengths of shells, do not happen so often as small deviations.

Similar results would be found with further examples, and we may conclude that shells possess a mid-length (or median) which is constant in different samples; but we are not yet justified in saying that other objects would give a similar result, and we must therefore examine other collections. Let us take the length and breadth of nuts. We will examine two lots, the first containing 185 specimens and the second 47.* Figs. 6

* The reader can choose any other object when he has mastered the idea. He must pick at random, and then arrange the specimens in their order. He then sees which is the middle term, measures it, and notes that this is the median.

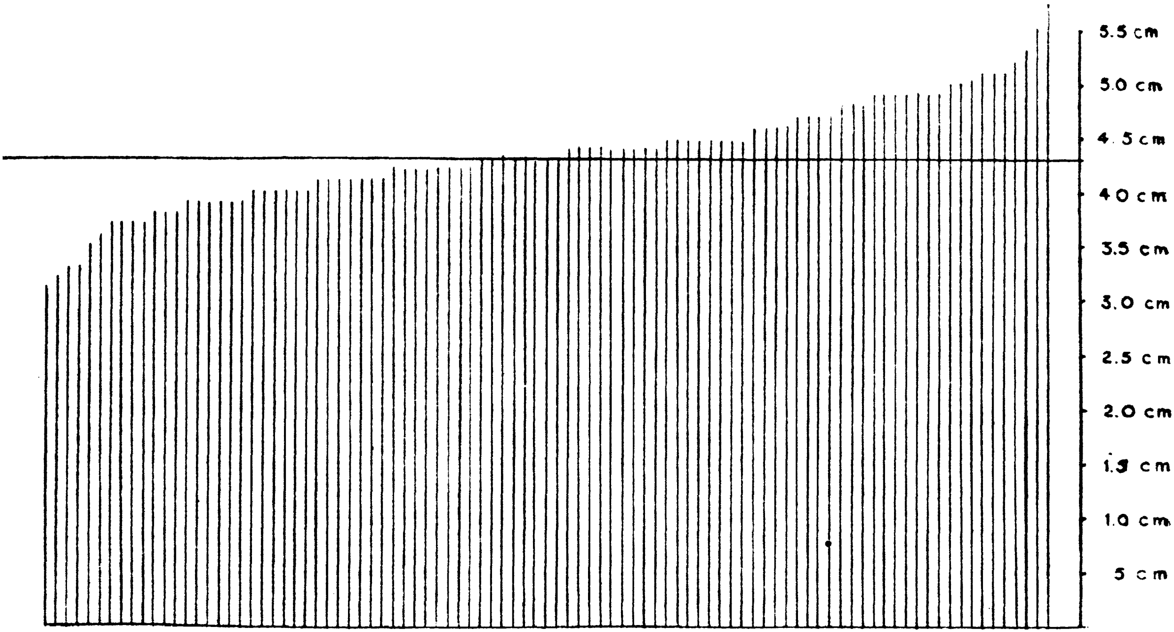

FIG. 6.—SHOWING LENGTHS OF 185 NUTS.
Only alternate nuts are shown to prevent the figure being unduly large.

to 9 give (1) the nuts in order of lengths, and (2) the nuts in order of breadths, for each collection. As a further example, the illustration facing page 1 is of interest; it is taken from a photograph of a number of pea-pods arranged by Sir Francis Galton,

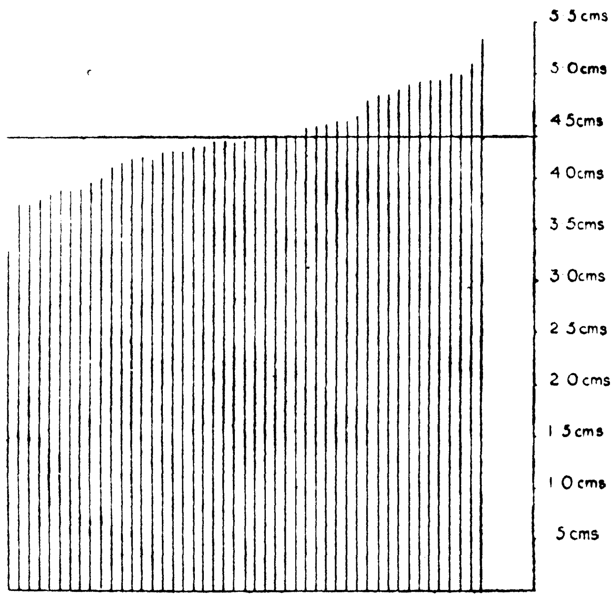

FIG. 7.—SHOWING LENGTHS OF 47 NUTS.

and instead of giving the lengths of each pod in a diagram the actual pods are shown. This could have been done with our measurements, but in most cases the objects would take up so much space that the method would obscure the shape of the curve. You will notice a considerable likeness

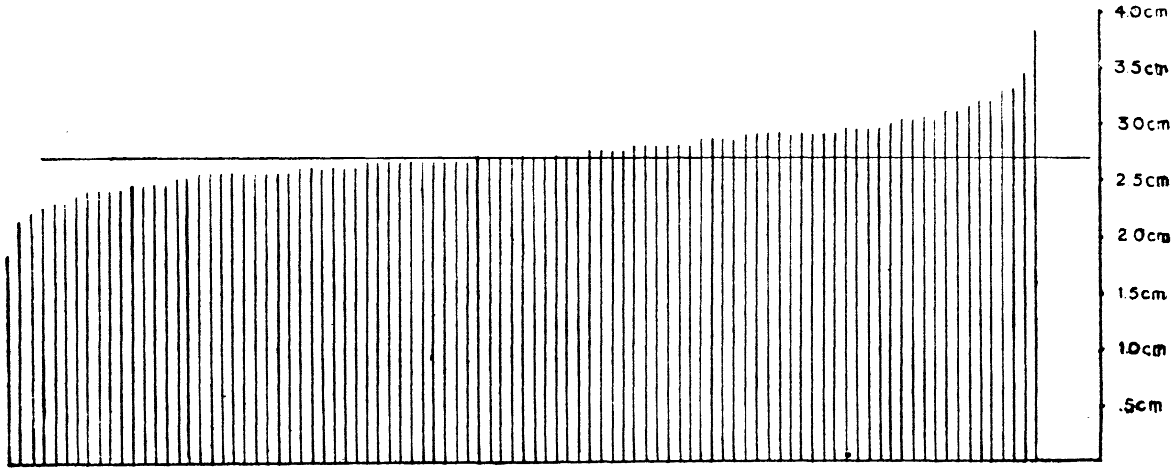

FIG. 8.—SHOWING BREADTHS OF 185 NUTS.
Only alternate are shown, as in Fig. 6.

between the figures in this chapter, whether they refer to nuts, shells, or pea-pods, and, in fact, the similarities in the shapes of the diagrams are so great that one feels there must be some general law behind all arrays, which would simplify our ideas of the sizes of things, and enable us to give

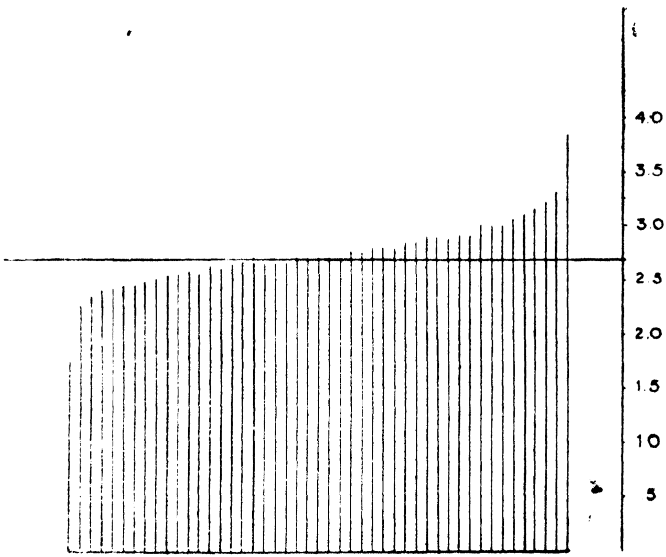

FIG. 9.—SHOWING BREADTHS OF 47 NUTS.

a fairly accurate idea of the size of an object capable of variation.

It is, however, advisable to go a step further in our inquiries. So far we have dealt with things which to the average person may imply a law, but we shall now touch on what is, in common parlance,

'sheer chance.' The following is the result of an experiment in coin-tossing: Fourteen coins were tossed 150 times, and the number of 'heads' was recorded each time. The cases were, in the order in which they occurred, as follows: 7, 6, 9, 7, 3, 6, 8, 6, 5, 11, 11, 6, 4, 7, 8, 7, 6, 6, 4, 6, 9, 7, 5, 4, 4, 3, 6, 11, 8, 8, 8, 7, 6, 6, 6, 5, 10, 7, 7, 8, 9, 8, 9, 7, 8, 4, 8, 9, 6, 4, 10, 8, 6, 7, 6, 5, 4, 11, 5, 6, 7, 6, 6, 4, 6, 9, 7, 6, 6, 9, 6, 4, 7, 6, 7, 9, 13, 4, 7, 7, 6, 8, 4, 5, 9, 5, 4, 10, 8, 7, 9, 8, 5, 7, 9, 7, 11, 8, 5, 5, 4, 4, 7, 8, 5, 4, 7, 5, 9, 8, 7, 6, 7, 5, 5, 7, 7, 7, 7, 6, 8, 8, 7, 9, 7, 10, 8, 9, 8, 8, 6, 7, 5, 7, 11, 10, 7, 4, 8, 7, 5, 10, 9, 7, 4, 7, 7, 5, 8, 8, 6.

This is a random order, like that given by the nuts when they were turned out of the bag, and if we arrange the tossings in order, as we did the nuts and shells, and draw vertical lines proportional to the number of 'heads' in each tossing, we get Fig. 10.

The form of the curve in this figure is similar to those already found, and suggests that if a theory can be evolved which explains these results of 'sheer chance,' it would probably assist us materially in dealing with our other cases. Many similar experiments in objects like shells and in matters of pure chance can easily bo made, and they will be found to concur in showing the same main features as those we have already examined; and

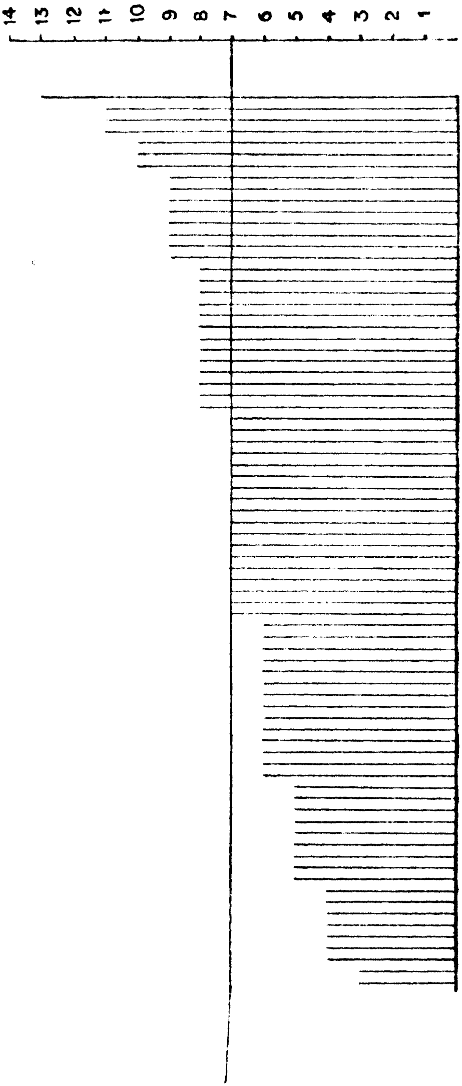

FIG. 10.—SHOWING RESULTS OF 156 TOSSINGS OF 14 COINS.

Only alternate tossings are shown to prevent the figure being unduly large.

we are led to conclude (1) that we get approximately the same median from various samples of the same object, and so the median is a useful measure of an array of variates; (2) that when we represent our work in a figure, we get approximately the same shape, whether we deal with nuts, shells, or coin-tossing, for it takes the form of a curve which rises rapidly at first, then keeps level, and finally rises rapidly at the end, showing, of course, that there are many cases nearly alike at the middle of the array (*i.e.*, near the median), and very few at each end (*i.e.*, differing much from the median). We would, however, add that although the form of the curve is very general, it is not quite universal, as will be seen in a later chapter.

CHAPTER II

QUARTILES AND MEANS

We have seen in the previous chapter that when a diagram is drawn from the measurements of different samples of certain classes of objects, the median remains about the same, and may be considered as a constant (*i.e.*, an unchanging value), and the curve at the top of the figure represents the variations of the object we have been examining. It is easiest to see these variations clearly by drawing a horizontal line through the top of the median (as has been done in all the figures of Chapter I.), and one cannot then fail to notice how this line brings out the comparative flatness of the curve near the median and the slopes at the two ends. This was referred to in Chapter I., but it is well to accentuate the fact that in all the cases we have dealt with there are only a few cases that show large deviations in size from the median, and many more cases that show small deviations—that is, are of nearly the same size as the median. We could express this shortly by saying that large

PRIMER OF STATISTICS

BY THE SAME AUTHORS

FREQUENCY CURVES AND CORRELATION. By W. PALIN ELDERTON.

Published by C. & E. LAYTON, 1906

NATURE AND NURTURE. By ETHEL M. ELDERTON.

Published by DULAU & CO., 1909

AGENTS

America . THE MACMILLAN COMPANY
64 & 66 Fifth Avenue, NEW YORK

Australasia. OXFORD UNIVERSITY PRESS
205 Flinders Lane, MELBOURNE

Canada . . THE MACMILLAN COMPANY OF CANADA, LTD
27 Richmond Street West, TORONTO

India . . . MACMILLAN & COMPANY, LTD.
Macmillan Building, BOMBAY
309 Bow Bazaar Street, CALCUTTA

We could easily continue the idea involved in the calculation of quartiles, and split the whole distribution into a hundred parts (percentiles), but when the number of objects measured is large, it is generally advisable to adopt a modified method of treating them, which leads to the more elaborate methods used in statistical analysis. We shall deal with these arrangements a little later, but before coming to them it is necessary to consider some of our collections a little more closely, and find what further can be learnt from them.

This brings us to the consideration of the *mean* of an array of variates. Everyone knows the meaning of the word 'average,' and the way it is calculated. If we want the average age of boys in a form, we add all the ages together and divide by the number of boys; or if we want to find a batting average, we add all the runs together and divide by the number of completed innings. In just the same way, if we want the mean length of a sample of shells, we add all the lengths together and divide by the number of shells.

The next step in our investigations will be to compare the sizes of means and medians. We will begin with a case which will be more familiar than shell measurements to many readers, and will take as an example the cricket scores of Tunnicliffe in 1907. These scores were abstracted from Wisden's

'Cricketers' Almanac' in order of date, and due allowance was made for 'not-out' innings by adding the 'not-out' score to that obtained in the following innings, and treating the two combined as one completed innings. This gives the same average as the usual method, and seems to be the idea underlying it. The only objection to it is that the conditions and opposing sides are not the same. We found that Tunnicliffe's mean score (average) was 30·2 runs, and if we arranged his scores in order of size (see Fig. 12), the median score was 30 runs. Put verbally, this would be interpreted as follows: In 1907 Tunnicliffe was as likely to make more than 30 runs as he was to make less than 30, and on the average he made 30·2 in each innings. The reader will notice that the mean and median do not differ much, but he will also see from the diagram that, owing to the large variations in the sizes of the scores made in different innings, the diagram looks rough, and its shape is not of the same kind as that with which we have been dealing up to the present. This is because a batsman is very likely to get out with quite a small score before he is used to the bowling and conditions. The diagram shows this because, if a curve were drawn through the tops of the ordinates, it would have to begin nearly horizontally; and we noticed, when dealing with the median in Chapter I., that when a curve is

horizontal, it means that we have a number of cases of about the same size. We have tried the scores of other well-known cricketers, and have found the same main characteristics, and we shall see later how such arrays can be dealt with fully.

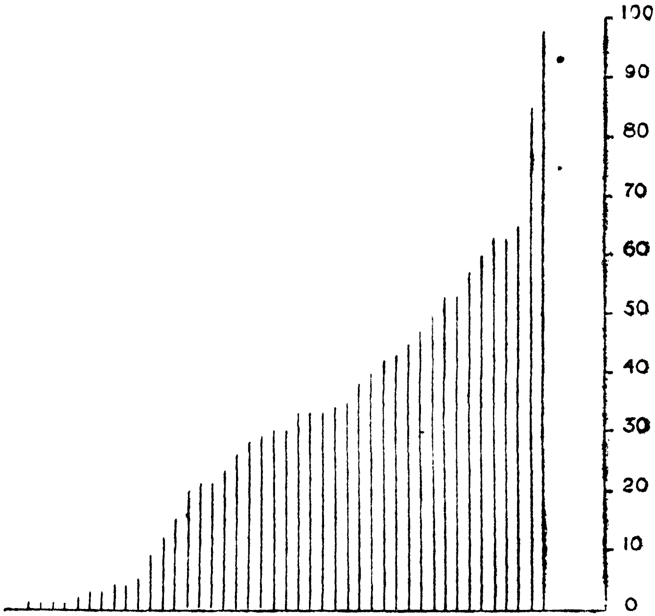

FIG. 12.—SHOWING TUNNICLIFFE'S SCORES, 1907.

For the present we are only concerned with the relative sizes of the means and medians, and we have made up the following table, giving the values of the means and medians for each case that we have considered up to the present, and we have also

included a case of the mean and median age obtained from 100 schoolboys in four forms at the Merchant Taylors School for the Christmas term, 1899.

TABLE I.

	Mean.	Median.
47 nuts (length) -	4·40 cm.	4·38 cm.
25 nuts (length) -	4·24 cm.	4·24 cm.
33 shells (length) -	3·20 cm.	3·20 cm.
33 shells (breadth)	1·53 cm.	1·61 cm.
59 shells (length) -	3·03 cm.	3·23 cm.
59 shells (breadth)	1·52 cm.	1·60 cm.
41 shells (length) -	3·21 cm.	3·21 cm.
41 shells (breadth)	1·58 cm.	1·56 cm.
100 shells (length) -	3·10 cm.	3·20 cm.
100 shells (breadth)	1·55 cm.	1·59 cm.
Tunnicliffe (cricket scores) - -	30·2 runs	30 runs
Ages of schoolboys -	13 yrs. 4 mos.	13 yrs. 6 mos.

It will be seen from the above table that the mean and median do not differ much from one another, and we will now consider a little more in detail the conditions which are necessary for the two values to be exactly equal. It is easiest to do this with the help of the following diagram, which is similar to those we have been using (Fig. 13). OM is the median, AMB the curved line showing variation,

and CMD the horizontal line through the top of the median. Now let us consider what would be the value of the mean of this array of variates. We add all the variates (vertical lengths), and divide by the number of them—that is, we balance the larger pieces after the mean with the smaller pieces before it. This shows us one case in which

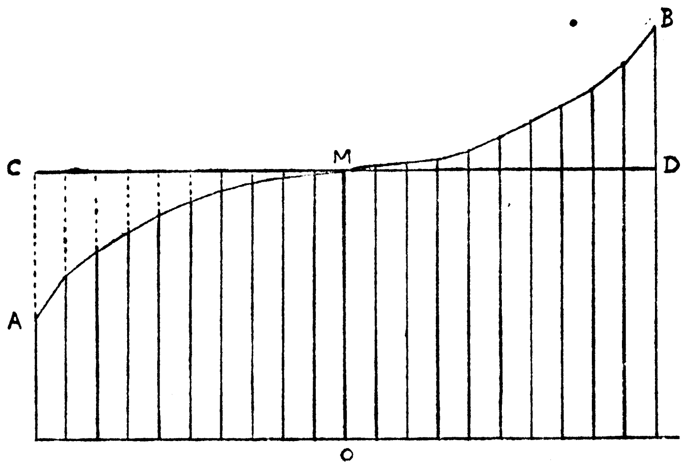

FIG. 13.

the mean and median are equal—namely, when the variate next after the median (or mean) is exactly as much larger than the median as the variate next below the median is smaller than the median, and when all similar cases show the same balance. In this case variation would be said to be 'symmetrical,' as each term balances exactly with a corresponding term. Of course, we can have the

mean and median equal in other cases, and the condition that they will be equal must be that the sum of all the pieces between the line MD and the curve MB will be equal to the sum of all the dotted lines which we have inserted between the curve AM and the line CM.

We can now summarize the work of this chapter. We have seen in the first chapter that large variations generally occur less frequently than small variations, and we have now seen that we can describe the rapidity of change in the size of adjacent variates, and give further information about the variations by recording other values beside the median, such as the quartiles. We also learnt how to calculate a mean, and found that it was not very different in value from the median in the cases examined; and, finally, we tried to see the conditions which would prevail if the mean and median have the same value. Briefly, the median expresses the average size, and the quartile expresses the spread or scattering of values.

CHAPTER III

FREQUENCY DISTRIBUTIONS

In our previous chapters we have in each example set out our measurements in order, by showing each individual case separately; but when we have collected a great number of specimens, it is easier to show their sizes by an alternative method.

So far, when we have been dealing with deviations we have reckoned them from the middle term (median), which is approximately equal to the mean in many cases, and we have seen that the curve at the top of the figure represents the entire system of deviations; but we could, instead of this, count the number of cases that fall in various groups on each side, either of the median or of the mean, and show that the number in each group becomes smaller as the group is farther from the mean.

By taking all the nuts or shells and setting them out in groups we can form the following series:

Example I.: Lengths of Nuts.

Number of nuts	2	7	28	59	49	33	6	1
Lengths in cm.	2·8–	3·2–	3·6–	4·0–	4·4–	4·8–	5·2–	5·6–

Example II.: Breadths of Nuts.

Number of nuts	2	0	5	22	55	48	36	10	5	1	1
Breadths in cm.	1·7–	1·9–	2·1–	2·3–	2·5–	2·7–	2·9–	3·1–	3·3–	3·5–	3·7–

Example III.: Lengths of Shells.

Number of shells	2	5	11	15	25	24	15	3
Lengths in cm.	0·9–	1·4–	1·9–	2·4–	2·9–	3·4–	3·9–	4·4–

Example IV.: Breadths of Shells.

Number of shells	3	9	11	14	23	23	9	6
Breadths in cm.	0·7–	0·9–	1·1–	1·3–	1·5–	1·7–	1·9–	2·1–

Example V.: Coin-Tossing (14 *Coins*).

Number of tossings	0	0	2	15	17	27	36	25	15	6	6	0	1	0
Number of heads	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Now, these series of numbers show at once that there are more individuals closely like the mean (average) than there are individuals who are less closely like it, but they show the fact in a rather different way from our examples in the previous chapters. Thus the first example says that out of 185 nuts there were 7 nuts which were between 3·2 centimetres and 3·6 centimetres in length, and 28 between 3·6 centimetres and 4·0 centimetres, and the process goes on till all the nuts are put in one or other of the groups. In other words, we have distributed the nuts in groups, and have, as it were, answered the question, 'If 185 nuts are collected, how are they distributed with regard to

length?' Or, 'If I take 185 nuts at random and distribute them in groups, how frequently shall I put one in each of these groups?' These statistical distributions are called 'frequency distributions.'

We will now draw this example as a picture in which each nut is represented by a square of the same size, so that the blocks are proportional in

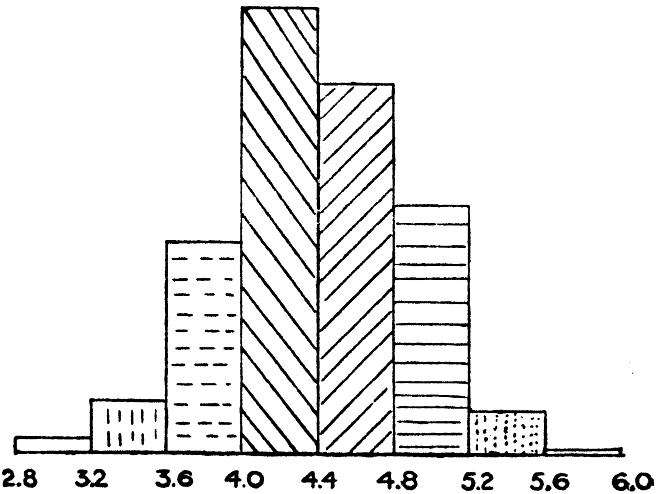

FIG. 14.

size to the number of cases. If the reader will compare this figure with the next one, he will see that the blocks shaded in the same ways correspond.

In Fig. 15 there are 49 cases which are between the mean and 4·8 centimetres, while in Fig. 14 the block is proportional to that number; we may say, in fact, that the heights of the blocks in Fig. 14

correspond to the breadths of the corresponding pieces in Fig. 15. This, however, is not very important, provided the reader clearly understands the two ways in which it is possible to deal with the statistics.

In the previous chapters we saw that we could draw a continuous line through the top of our

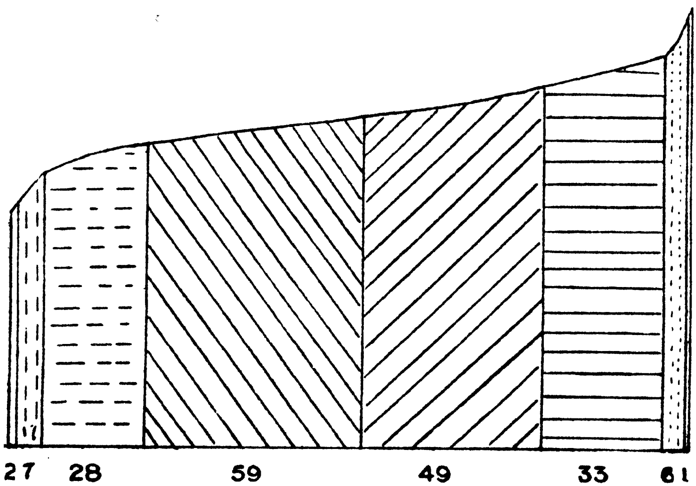

FIG. 15.

ordinates (vertical lines), and, by examining it, get an idea of the deviations of the object. In the same way we can draw a bell-shaped curve through our blocks, and study it instead of the rough polygon.

The consideration of these smooth 'frequency curves,' as they are called, leads one to the more

mathematical side of the subject; but it is quite possible to appreciate their importance, even though the mathematical work with which they are associated is neglected.

In order to do this, let us return for a while to the rough polygon (Fig. 14) which proceeds up and

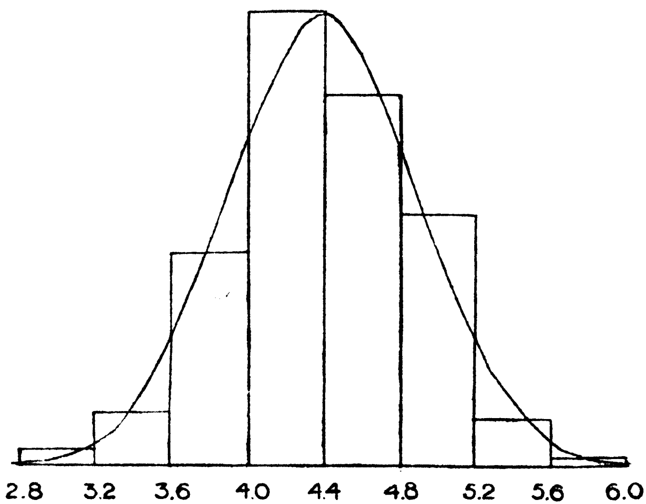

FIG. 16.

down by a series of steps. We have noticed that any group far from the mean will contain fewer cases than one near the mean, and this leads one to expect that if we were to subdivide any one of our existing groups into two parts, the part nearer the mean would contain more cases than the one farther off. Thus, if we take the group relating to

4·0 to 4·4 centimetres in length of nuts, we have 59 cases. If we divide it into two groups, 4·0 to 4·2 centimetres and 4·2 to 4·4 centimetres, the former contains 24 and the latter 35. If we divide the next term a similar thing happens, and we notice that if this is continued throughout the table the steps become smaller and more frequent. Now, if we repeat this division process the steps will again become smaller, and if we go on long enough we shall smooth them entirely away, and have smooth slopes instead of the steps we started with. If you try this you will, however, meet a difficulty. Starting with 185 cases in 8 groups, and breaking up each group into 2 parts, you will find the sub-groups will in some cases contain so few cases that they will not help you to see the way in which the statistics run; this means that you have not started with a sufficient number of cases. We might restate our difficulty by saying that we can only reach a perfectly smooth curve when we have an infinitely large number of cases. The smooth curve is, as it were, the ideal to which the statistics would lead if we had a sufficient number of cases.

An easy way of seeing how the curve and steps (polygon) correspond is to start with a large number of cases divided into many groups, and then re-arrange them in fewer groups. The series of

numbers in Table II. will enable the reader to do this. The first column gives the numbers that

TABLE II.

	Number of Cases.	Adding together in Pairs.	Adding previous Column.	
	1	3		
	2		12	
	3	9		
	6			
	12	33		
	21		117	
	34	84		
	50			
	69	156		
	87		371	
	103	215		
	112			
Mean				Mean
	112	215		
	103		371	
	87	156		
	69			
	50	84		
	34		117	
	21	33		
	12			
	6	9		
	3		12	
	2	3		
	1			
	1,000	1,000	1,000	

arise in each small group out of 1,000 cases; in the next column we have added our groups together in pairs, so that we have applied the opposite method to that mentioned above, and if the reader wishes to follow that argument again he must begin at the right-hand column and work backwards to the left hand.

The following diagram (Fig. 17) will show how the steps gradually merge into the smooth curve. In order to save space we have only drawn one-half of the figure.

The reader will notice that the steps are artificial, and arise because we must make arbitrary limits in statistics, as by collecting them in inches, or years of age, or pounds; but as such divisions are purely arbitrary (for we might use half inches, or months, or ounces), we need to replace them by something which is not connected with such arbitrary arrangements. The smooth curve is independent of groups, and is therefore of a more general nature than the rougher polygon.

It will be interesting to mention here a use to which these frequency curves can be put. If we refer to Fig. 16, and imagine that we merely know the rough figures represented by the blocks—*i.e.*, we only know the frequency polygon—it is conceivable that in some cases we might want to know the number that would fall in some other group—say

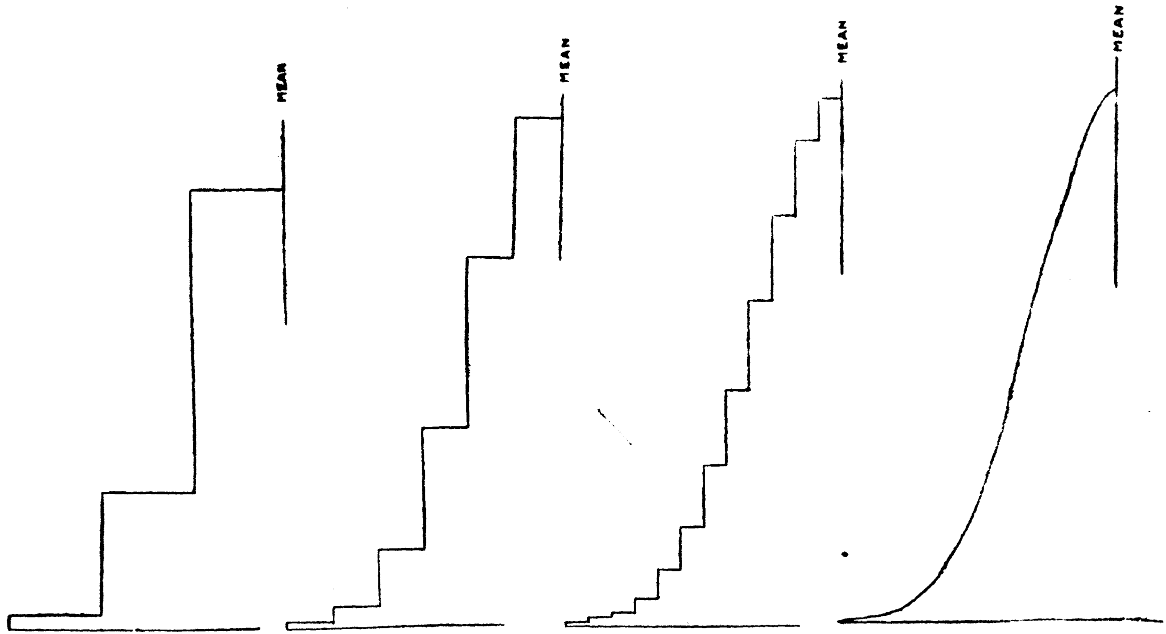

Fig. 17.—Showing how Statistics lead to a Smooth Curve when Finer Grouping is Adopted.

from 4·2 to 4·6 centimetres. This might easily be wanted in connection with census returns, as such statistics are frequently obtained only for groups of ten ages, say 30 to 40, 40 to 50, etc., and we might for some purpose require to know the number between the ages 32 and 49. Now, assuming that such a curve has been drawn in the present instance, in order to find out how many cases there would be in the new group, we have only to measure the size of the figure bounded by the curve at the top, the base from 4·2 to 4·6 centimetres, and vertical lines drawn from the ends of this base to the curve. The reader may now ask why one does not in such circumstances count up the number of cases from the original statistics. The reason is that it is often impossible, as it would have been in the census case mentioned above.

Let us now consider another purpose to which our frequency curves can be put. Take any one of the examples at the beginning of this chapter, and try to see what would have happened if we had only dealt with very few cases. Suppose we take 12 shells at random; then we find, on arranging them, that the groups corresponding to those of Example I. are as follows:

Group - - -	2·8–	3·2–	3·6–	4·0–	4·4–	4·8–	5·2–	5·6–
Twelve specimens	0	1	2	1	4	3	1	0
Example I. - -	2	7	28	59	49	33	6	1

We see that there is a similarity between the two series, but the 12 specimens run very unevenly in comparison with those of Example I., while that example itself does not give a perfectly even series. Now, remember that the object of measuring is to find the sizes of all the shells that could possibly be collected. But the more shells we collect the more even will our series become, till ultimately it is absolutely smooth. If, therefore, we can by some method draw correctly—through either the 12 specimens or the 100 cases in Example I.—a smooth curve, we shall have removed the roughnesses due to the use of only a few cases, and got nearer to the smooth ideal result that would be obtained if we could measure every shell in existence. Great care is, however, necessary; a curve drawn freehand through the statistics may be far from giving the most correct estimate possible of the ideal result, and so mathematical methods have to be adopted. The curves are to be chosen on two grounds: (1) By considering what happens when we deal with problems in pure chance, like tossing coins or throwing dice; and (2) by seeing, from the examination of a large number of various kinds of statistics, the shapes of the curves that generally occur. We do not intend to go deeply into this part of the subject, but we will endeavour to give a slight idea of the connec-

tion between chance and the smooth curves we have mentioned. For this purpose we will take the simplest case, and imagine that we toss four coins, and write down all the possible results that can occur. They are as follows, where an 'h' is written for 'head' and a 't' for 'tail':

First Coin.	Second Coin.	Third Coin.	Fourth Coin.	
h	h	h	h	1 case of 4 heads.
h	t	h	h	
h	h	t	h	4 cases of 3 heads.
h	h	h	t	
t	h	h	h	
h	h	t	t	
h	t	h	t	
h	t	t	h	6 cases of 2 heads.
t	h	h	t	
t	h	t	h	
t	t	h	h	
t	t	t	h	
t	t	h	t	4 cases of 1 head.
t	h	t	t	
h	t	t	t	
t	t	t	t	1 case of 0 head.

The following table, which is exactly similar to those of the examples at the beginning of the chapter, gives the same result:

TABLE III.

Number of 'Heads turning up.	Number of Cases.
0	1
1	4
2	6
3	4
4	1
Total	16

This is the most likely result. It is, however, not certain that if 16 tosses were made such a distribution would be realized, as there are many other possible results, and some of them are nearly as probable as the one we have given. Now, if we made 16 tosses, and found 1, 3, 7, 4, 1, as the result, instead of 1, 4, 6, 4, 1, then we could say that the former was the result of the 'statistics' we had collected; but if we went on for a very great number of times, the proportions we should get would approximate more and more closely to those of 1, 4, 6, 4, 1. Now, there is no special reason for choosing 4 coins, and if we had taken 8 we should have found the following series :*

* To those with any experience of algebra it will be at once apparent that the terms in the series of heads are given by the terms in the expansion of $(1+1)^n$, while those who have not such

Number of heads	0	1	2	3	4	5	6	7	8
Number of cases	1	8	28	56	70	56	28	8	1

If this is drawn in a figure, you will notice that the smooth curve drawn through it is of a bell shape, like that already noticed (Fig. 16), and that it is symmetrical about the mean—that is, after the mean the first half is repeated in opposite order. You will also notice that the middle terms are argest, and at the end they are small, and as we deal with more coins you will find, if you draw figures like Fig. 16, that the 'tails' of the curve run close to the horizontal line at the bottom of the figure. A certain curve, which can be evolved like this, is known as the 'normal curve of error ; the 'Gaussian curve,' from its discoverer; or sometimes also as the 'probability curve,' though the last is a misnomer, as there are other curves that express the probabilities of the happening of events.

---

acquaintance can always obtain the terms by the method implied in the following table, in which any term—say the sixth in any column—is the sum of all the previous terms in the previous columns, or, more easily, is the sum of the fifth term in the column and the sixth in the previous column.

		1								
		1	1							
2 coins	- -	1	2	1						
3 ,,	- -	1	3	3	1					
4 ,,	- -	1	4	6	4	1				
5 ,,	- -	1	5	10	10	5	1			
6 ,,	- -	1	6	15	20	15	6	1		
7 ,,	- -	1	7	21	35	35	21	7	1	
8 ,,	- -	1	8	28	56	70	56	28	8	1

These other curves, many of which are not symmetrical, can be evolved from similar considerations. Thus if 6,561 throws are made with 8 dice, and a record is kept of the numbers of 'fives' and 'sixes' that turn up, the most likely numbers would be those shown in Table IV.,* which the reader who is unacquainted with the theory of probability will have to take on trust.

TABLE IV.

Number of 'Fives' and 'Sixes.'	Number of Times the Number mentioned in the Previous Column turned up.
0	256
1	1,024
2	1,792
3	1,792
4	1,120
5	448
6	112
7	16
8	1
Total	6,561

* It will be of help to those knowing algebra to bear in mind that the series is obtained from the expansion of $(1+2)^8$, the chance of getting a 'five' or 'six' being half the chance of not getting either.

It will be noticed that it is only on the average once in 6,561 times that every die thrown turns up a 'five' or 'six.'

Another curve showing a still greater divergence from symmetry can be reached by considering the number of times sequences occur. Since it is equally likely if we toss a coin for it to come down 'heads' or 'tails,' we are half as likely to get a sequence of two (*i.e.*, two heads followed by a tail, or two tails followed by a head), as we are to get a sequence of one, and also half as likely to get a sequence of three as we are to get a sequence of two, and so on, as shown in Table V.

TABLE V.

Sequence of—	Number of Times Sequence occurs.
1 head or tail	16
2 heads or tails	8
3 ,, ,,	4
4 ,, ,,	2
5 ,, ,,	1

The curves evolved from these and similar probabilities have been found to describe statistics of very various classes, and it is therefore reasonable to use them for approximating to the ideal result

to which we have so often referred. In any individual case arithmetical calculations are required of a rather lengthy description, but it is unnecessary to go into them, as we only wish to indicate how such curves arise and why they are wanted.

We may now summarize our work on them :

1. They remove the steps that occur, owing to statistics being collected in inches or some similar arbitrary measure.

2. They help us to estimate the ideal result, which we might reach if we could use a large number of cases.

3. By starting from results that can be obtained by general reasoning, such as coin-tossing, we can form an idea of the types of curves we should generally obtain.

CHAPTER IV

MODE—STANDARD DEVIATION—COEFFICIENT OF VARIATION

In Chapter II. we gave as an example the number of runs made in an innings during 1907 by Tunnicliffe, and found that he was as likely to make more than 30 as he was to make less than 30, and on the average he made 30·2 in each innings (*i.e.*, the median was 30 and the mean 30·2).

Let us suggest two more problems, with which we will deal in order:

1. What was Tunnicliffe's most likely score in 1907?
2. Was he a consistent batsman?

1. At first sight it might seem that the average number of runs a man makes would be that which is his most likely score in any particular innings. A little consideration will show that this is not necessarily the case. Imagine the case of a cricketer who seldom made runs at all, and assume that he made the following scores in succession:

5, 6, 9, 8, 5, 0, 1, 3, 12, 0, 2, 0, 5, 0, 11, 8, 0, 10, 0, 1, 2, 6, 1, 0, 3;

or 25 innings for 98 runs, or an average of 3·92; but if you count up the number of innings and arrange them as a distribution you will find:

Number of runs in innings	0	1	2	3	5	6	8	9	10	11	12
Number of innings	7	3	2	2	3	2	2	1	1	1	1

From this table you will see that this batsman never made 4, and only twice made 3 in an innings; but on 7 occasions he did not get any runs at all, and 3 times he only made 1 run; while he made less than his average on 14 occasions, and more than his average on only 11. Clearly if you had to guess how many runs he made in some particular innings you would say he did not get any, and in 7 cases out of 25 you would be right. If you guessed 4, you would be wrong every time. This shows that the 'average event' is not always the 'mode,' which is the technical term for the most likely event. You will notice that the median in this case is a little nearer the mode than the average. It will be found that the median is given by the thirteenth term, which implies 3 runs. Now turn to Tunnicliffe's scores, and you will notice that even with such a good batsman there were many small innings. This is easy to see for yourself,

and the following groups of scores help to show it clearly; they were based on the three years 1905, 1906, and 1907:

Number of runs	0–9	10–19	20–29	30–39	40–49	50–59	60–69
Number of innings	39	25	19	15	7	11	5

Number of runs	70–79	80–89	90–99	100–109	110–119	120–129	130–139
Number of innings	4	3	3	1	2	1	2

If you were asked to guess in which group a particular innings fell, you would say the first; in other words, you would assert that, though Tunnicliffe's average was well over 20, he was most likely to make under 10 in a particular innings.

The reader should try other examples. In order to assist matters for those who are interested in cricket averages we have tabled the scores of a few well-known cricketers over the three seasons 1905, 1906, and 1907, and they are shown below. The scores are arrayed in the order in which they were made, but 'not-out' innings are added to the next innings as if it was mere continuation, as was done in Chapter II.

Hayward.

1905 43, 27, 13, 9, 10, 34, 59, 53, 13, 116, 11, 4, 3, 31, 22, 35, 3, 24, 21, 128, 168, 52, 122, 17, 148, 91, 88, 14, 203, 13, 64, 177, 88, 76, 106, 112, 2, 25, 48, 64, 26, 24, 216, 81, 58, 14, 33, 32, 35, 53, 10, 9, 197, 12, 28, 28, 0, 44, 2.

1906: 9, 37, 24, 24, 58, 10, 22, 255, 13, 73, 1, 5, 47, 3, 110, 80, 0, 16, 8, 21, 20, 3, 31, 70, 55, 26, 60, 68, 32, 143, 80, 5, 19, 1, 115, 82, 70, 52, 6, 85, 6, 33, 31, 6, 59, 2, 66, 71, 63, 44, 8, 40, 122, 29, 38, 76, 32, 17.

1907: 39, 82, 25, 5, 219, 135, 0, 46, 0, 68, 10, 31, 194, 244, 143, 125, 54, 69, 6, 15, 100, 144, 34, 144, 208, 54, 34, 76, 10, 16, 10, 109, 51, 18, 14, 76, 168, 17, 3, 94, 45, 51, 8, 9, 24, 60, 110, 141, 22, 28, 62, 51, 17.

Jessop.

1905: 0, 0, 4, 20, 61, 26, 8, 87, 10, 29, 10, 52, 5, 27, 40, 12, 2, 11, 34, 206, 48, 43, 0, 1, 1, 3, 7, 27, 77, 17, 43, 42, 24, 20, 23, 10, 68, 170, 58, 47, 77.

1906: 28, 1, 60, 38, 14, 48, 0, 0, 1, 25, 1, 17, 20, 11, 14, 89, 234, 5, 23, 0, 1, 52, 4, 2, 1, 4, 19, 15, 54, 21, 65, 57, 60, 8, 10, 15, 11, 1, 45, 0, 74, 8, 25, 0, 43, 0, 1.

1907: 63, 55, 5, 17, 16, 42, 8, 8, 87, 27, 22, 34, 4, 48, 5, 80, 66, 40, 0, 15, 6, 12, 111, 0, 4, 19, 53, 4, 6, 0, 48, 15, 75, 26, 26, 28, 1, 10, 34, 2, 14, 9, 54.

Tunnicliffe.

1905: 15, 85, 12, 77, 9, 18, 12, 9, 16, 4, 24, 8, 119, 2, 18, 6, 24, 20, 73, 50, 15, 94, 31, 0, 0, 128, 10, 135, 0, 139, 15, 40, 55, 10, 64, 26, 18, 20, 32, 34, 0, 0, 90, 13, 80, 18, 16, 28.

1906: 3, 71, 12, 0, 9, 43, 0, 52, 2, 102, 0, 117, 24, 6, 15, 1, 71, 3, 5, 24, 0, 16, 4, 31, 22, 3, 50, 36, 58, 37, 17, 9, 6, 13, 10, 11, 33, 59, 20, 57, 25, 14, 16, 29.

1907: 65, 5, 2, 98, 4, 26, 9, 63, 29, 30, 33, 40, 1, 21, 45, 33, 20, 50, 3, 0, 0, 42, 35, 47, 1, 12, 60, 53, 34, 15, 43, 57, 1, 63, 1, 33, 23, 3, 4, 38, 28, 30, 53, 21, 85.

Warner.

1905: 27, 76, 14, 13, 47, 45, 39, 7, 15, 34, 38, 106, 107, 75, 8, 3, 4, 4, 4, 47, 13, 209, 66, 14, 0, 78, 10, 23, 0, 0, 44, 16.

1906: 0, 10, 204, 0, 0, 60, 0, 85, 0, 49, 86, 3, 50, 5, 10, 7, 70, 43, 9, 59, 97, 0, 166, 152, 13, 22, 34, 82, 14, 14, 35, 56, 48, 23, 31.

1907: 61, 42, 137, 10, 41, 72, 0, 22, 53, 66, 4, 73, 28, 17, 16, 122, 48, 5, 87, 26, 9, 66, 7, 9, 17, 44, 60, 12, 66, 2, 77.

The same main characteristics will be found in each case, and we are forced to the conclusion that the average does not measure the most probable score.

Now turn to shells or nuts. What is the most likely length or breadth? The problem is the same as that of the cricket scores; but here it seems as if the most likely size is near the mean. What is the explanation? How can we get a measure of the most probable length or score (mode) since it seems to be a very wandering element?

We will try to explain a method by means of the curves we were discussing in the last chapter.

In drawing the distributions we arrayed our statistics in groups, and said that we would put in

the same group all the cases that were of approximately the same size. Now, if there are more of any one size than of any other, the curve we draw through our statistics will have to be tallest at that particular size which is 'most likely.' Now compare the two curves we have drawn. In Fig. 19 the curve is symmetrical; in Fig. 18 it clearly is not.

Now we saw in Chapter II. (p. 21) that when a

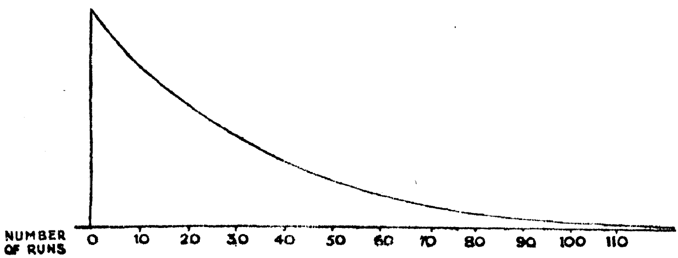

FIG. 18.—CURVE FOR CRICKET SCORES, SHOWING THE NUMBER OF TIMES EACH SCORE IS MADE, THE SCORE MADE MOST FREQUENTLY BEING 0.

curve is symmetrical the mean and the median are the same, and the curve is the same each side of the mean; so it is not hard to see that the mode will also have to be the same as the mean whenever the curve is symmetrical.

In the curve in Fig. 18 the mode is to the left of both the mean and the median, and if you draw a number of curves, from various series of statistics, and measure the difference between the mean and mode, and between the mean and the median, you

will find that the former is about three times the latter, and that the median lies between the mean and the mode. This gives a good approximation in many cases, but you will find that it will break down sometimes in extreme cases, such as the cricket scores. When this happens we have to use more

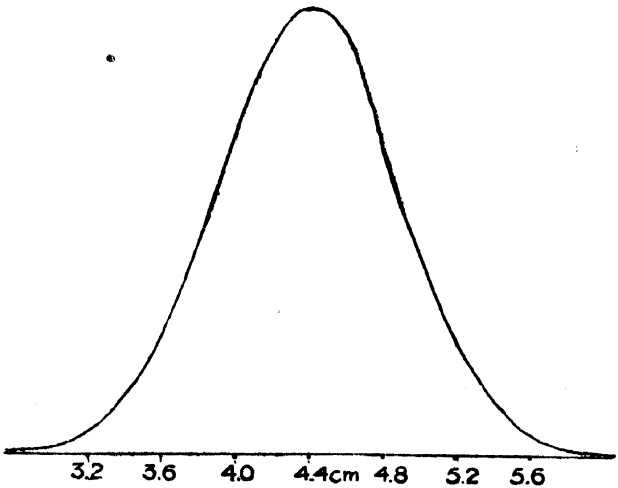

FIG. 19.—CURVE FOR NUTS, SHOWING THE NUMBER OF NUTS OF EACH LENGTH, THE GREATEST NUMBER OF NUTS BEING 4·4 CM. LONG.

troublesome methods; but you will find you can often get a rough idea of where the mode is in such extreme cases by drawing a curve through the statistics; and if you are not too dogmatic about your result, and compare it with that obtained by others by more refined methods, you will soon get to estimate the mode fairly well with a little practice.

2. Now let us turn to our second question, and try to see how we can find out if the batsmen we have mentioned were consistent in scoring.

In order to answer this question we must see what meaning we intend to attach to consistency, and in what direction we must proceed to get an idea of its value. If you compare two batsmen, you may express the opinion that one, say Hayward, is far more consistent than Jessop, and your reason for expressing the opinion is probably that you have noticed Hayward is less likely to make a very small score than Jessop. Now it appears, as soon as one begins to think it over carefully, that this of itself is hardly sufficient evidence, for a man may be unfortunately consistent in making small scores. If a man always made 25 runs whenever he went in to bat, and never made more or less, we should all agree that he was absolutely consistent; while if he generally made 25 runs, and sometimes a few more and sometimes a few less, we should say he was fairly consistent. In other words, we should say that, if it is desired to study consistency, it is necessary to examine deviations, for, it seems, the wider the deviations, the less the consistency. We can, however, see that this is only part of what is required, for if one player made 24, 25, 26, and another 34, 35, 36, it is reasonable to say that the latter is more consistently near his average score

than the former, because the second man had a wider chance of missing his average. He might have got anywhere between 0 and 34, while the first man only had the chance of being inconsistent up to 24; and the man with the lower average cannot claim that he could have been more inconsistent in the direction of a heavy score, for a deduction of 10 from infinity is unimportant, even though an 'infinite' score is not a thing likely to occur in practice! Again, if two men made 4, 6, 8 and 8, 12, 16, they should be equally consistent, because, if we imagine that the first is like a man shooting at 6, and the latter like one shooting at 12, we can see that a deviation of 2 in the former case corresponds exactly to a deviation of 4 in the latter. When these ideas are summed up, we conclude that we must consider not only the deviations themselves, but their relations to the mean from which the deviations are measured, in order to estimate consistency properly.

We will discuss the deviations alone first, and come to the measure of consistency afterwards, and for this purpose will turn back for a moment to Chapter II., and recall the quartiles we calculated. The reader will remember that they represent in a way the amount of deviation. Just as the median gave us an idea of the size of a thing, so the quartile gave us an idea of its tendency to deviation.

It would seem, therefore, that medians and quartiles might be employed; but though they are sometimes useful, they are open to objection. If you take almost any example, and find its quartiles, you will see that the one above the median will not be quite the same distance from the median as the one below; they will, in fact, be only equidistant from the median if the statistics are symmetrical. Now, if we found that for some batsman his median score was 35, and the lower quartile was at 10 and the upper at 80, we should merely be able to give these three figures, and it would be very difficult to say whether he was more or less consistent than another batsman whose scores gave a median of 37 and quartiles at 15 and 90. The fact that the two quartiles are not always the same distance from the median is an objection to their use in many cases; and another objection to quartiles is that they are reckoned with reference to the median and not the mean, while the latter is more frequent in practice, and may be distinctly different from the former. It must not, however, be inferred that the quartiles are not useful. They are a great help in forming a rough and rapid opinion of values, and they can often save one the trouble of calculating values by more lengthy methods. Statisticians have got over these difficulties in using them by adopting *a sort of* average measure of deviation, which is called the

standard deviation. This function treats deviations above and below the mean on the same terms, and does not get rid of the fact that the curve is not symmetrical, but it obscures the difficulties already referred to, and gives a definite result on which to work. To find it we may start from the average score and work in the way shown in Table VI., which is based on Warner's scores in 1907 His average was 41·9.

The first column is taken from p. 44; the second shows how much above or below the average any particular score was, the scores above the average having + in front, and those below having −. Now, if you add up the + part of this column, you will find it is equal to the − part, which should, of course, be the case, because of the way an average is calculated. The third column is found by squaring the second, and the rest of the calculation is shown afterwards. The standard deviation can be obtained roughly by adding the distance of the two quartiles from the median together and multiplying the result by 0·7414.* In the cricket scores of Warner this would have given a standard deviation of 38 (true standard deviation 35), but it is a case where this rough rule (which depends to some extent on the symmetry of the material) could not be expected to give a very good result.

* This follows from mathematical consideration, discussed in all technical works on Probability.

TABLE VI.

Scores.	Deviations from Means.	Square of Deviations.
61	+ 19·1	364·81
42	+ 0·1	0·01
137	+ 95·1	9044·01
10	− 31·9	1017·61
41	− 0·9	0·81
72	+ 30·1	906·01
0	− 41·9	1755·61
22	− 19·9	396·01
53	+ 11·1	123·21
66	+ 24·1	580·81
4	− 37·9	1436·41
73	+ 31·1	967·21
28	− 13·9	193·21
17	− 24·9	620·01
16	− 25·9	670·81
122	+ 80·1	6416·01
48	+ 6·1	37·21
5	− 36·9	1361·61
87	+ 45·1	2034·01
26	− 15·9	252·81
9	− 32·9	1082·41
66	+ 24·1	580·81
7	− 34·9	1218·01
9	− 32·9	1082·41
17	− 24·9	620·01
44	+ 2·1	4·41
60	+ 18·1	327·61
12	− 29·9	894·01
66	+ 24·1	580·81
2	− 39·9	1592·01
77	+ 35·1	1232·01
		37392·71

$\frac{37392{\cdot}71}{31} = 1{,}206$, average of squares of deviations.

$\sqrt{1{,}206} = 34{\cdot}73$, standard deviation.

We will now set out a table of the results we have found for the batsmen mentioned on pp. 42-44 for three separate years, and also for the three years taken together, which is of interest for comparison.

TABLE VII.

	Average.	Standard Deviation.	Coefficient of Variation.
Hayward:			
1905 - -	54·9	55·0	100
1906 - -	44·5	43·9	99
1907 - -	66·4	62·0	93
1905-1907 -	54·9	54·7	99
Jessop:			
1905 - -	35·4	42·1	120
1906 - -	26·1	38·5	148
1907 - -	27·9	27·0	97
1905-1907 -	29·6	36·7	124
Tunnicliffe:			
1905 - -	35·8	38·2	106
1906 - -	25·6	27·5	108
1907 - -	30·2	23·9	79
1905-1907 -	30·2	31·0	103
Warner:			
1905 - -	37·1	43·1	116
1906 - -	43·9	49·3	112
1907 - -	41·9	34·7	83
1905-1907 -	41·0	43·2	105

It will be seen at once from this table that not only was Hayward the highest scorer of those examined, but also the most consistent batsman, as the ratio of the standard deviation found from his scores to his average score (see p. 52) is lower than that of the others, though Tunnicliffe and Warner in 1907 were wonderfully consistent (see column headed 'Coefficient of Variation,' which is 100 × standard deviation ÷ mean). It will also be observed that all four batsmen were more consistent in 1907 than in the previous years, which may possibly be the result of the atmospheric conditions of that year.

We will now leave cricket scores, and interpret the same statistical functions for nuts and shells. It will be unnecessary to go into the preliminary explanations again, and the most useful way will be to show the results in the following table:

TABLE VIII.

	Mean in Centimetres.	Standard Deviation in Centimetres.	Coefficient of Variation.
100 Shells:			
Length - -	3·10	0·77	25
Breadth - -	1·55	0·34	22
185 Nuts:			
Length - -	4·40	0·54	12
Breadth - -	2·75	0·27	10

From these figures it will be clear that the shells and nuts vary far less than the cricket scores, and if the reader will think over Figs. 18 and 19, he will see that the reason is that in the case of the nuts there is a very small proportion lying far from the mean, while in the case of cricket scores there are a great many innings as far below the mean as they can be—*i.e.*, near 0. It will also be noticed that the nuts show less variation than the shells.

The reader can easily imagine problems in the solution of which such comparisons would be of assistance—as, for instance, in the studies of the measurements of skulls found in various districts—and he will see that, by the use of the statistical coefficients we have explained, one could go some distance towards deciding whether a certain set of skulls belonged to the same race as another.

CHAPTER V

CORRELATION

So far we have been confining attention to one thing at a time—thus we dealt with the lengths or breadths of nuts or shells, or with the scores of cricketers; but it is frequently more important in statistical work to deal with the relation of two things to each other than it is to examine the variations of one thing alone. Let us put this new problem in a concrete form: Is there any relation between the length and breadth of nuts or between the length and breadth of shells?

We must examine this question generally first, and see that we understand what it implies. Suppose we took a nut out of a collection of nuts, and found it was a long one: ought we, *from this information only*, to conclude it was a broad nut? If so, we are assuming that length and breadth are related, owing to some cause or other. If, on the other hand, we said it was impossible to estimate in any way the breadth of a nut from

knowing its length, we should be assuming that length and breadth in nuts were not related.

The same question can be asked about shells; and then, if we found that in both cases there was a relationship (correlation), we might ask whether there was a closer relationship between the lengths and breadths of nuts than between the lengths and breadths of shells, or *vice versa.* This suggests that what we must find is some way of comparing our correlations, so that, at the end of our work, we can arrange 'relationships' or 'correlations' in order, just as we have, up to the present time, been arranging lengths and breadths: as we fix a scale of inches (say) for measuring heights, so we must fix a scale of 'somethings' for measuring relationships.

If, whenever we took up a nut, we knew that its breadth was always exactly half that of its length, then we could say the lengths and breadths of nuts were absolutely related to one another; or if we knew that there was always a difference of 2 centimetres between the lengths and breadths, then we could say the relationship was absolute; in fact, whenever a fixed connection always holds between two things, we can say that they are absolutely related, or, in more technical language, that the correlation is perfect. This absolute relationship must come at the top of our scale, and statisticians, in making their

scale, use unity as representing perfect correlation, or state that the 'coefficient of correlation' is unity.

Now, suppose you knew that the length of a nut was about twice as great as the breadth, but sometimes it was a little more and sometimes a little less, then you have a relationship which is nearly, but not quite, absolute; you get something which is nearly perfect, but not quite. In these cases a value is required on the scale a little below that which we have used, and we should have to do some arithmetical work to see how far down the scale any particular case should come.

We will examine the case when there is no relationship at all. We know that the average breadth worked out from nuts of all sizes is B centimetres; now, if we found that, when we collected all the longest nuts, their average breadth was B centimetres; and if we also found that, when we collected nuts of medium length, their average breadth was also B; and again, if we took the short nuts, and found that their average breadth was B as well, then I think you will see that, if we were told the length, we should not be better able to estimate the breadth of the nut than if we were merely told that it was a nut. In such cases there is no relationship, therefore, between length and breadth, and we say that the 'coefficient of correlation' is zero.

So far, then, we have seen that we can express relationships on a scale which runs from 0 to 1; but you will find that, even so, we have not covered all possible cases that may arise, for we have merely imagined examples in which the long nuts were broad or of average length; but if we found that whenever a nut was long it was narrow, and whenever it was short it was broad, we should have a relationship between length and 'narrowness' exactly like that between length and breadth, with which we have already dealt. But, remembering that 'narrowness' is only breadth from the opposite point of view, statisticians merely extend their scale backwards to -1, and have therefore a scale of coefficients of correlation running from -1 to $+1$; but though it is convenient to have this negative part of the scale, it need not cause any difficulty, for if we find it troublesome we can always say we are talking about 'narrowness' instead of 'breadth,' and our coefficient will become positive.

We must now turn to the arithmetic of the question, and will concern ourselves first with the nuts, and for this purpose Table IX. may be taken; it explains itself fairly well, but it will probably be advisable to run through some of the figures. Thus, taking the third column, we see that there were 5 nuts, which were between 2·1 and 2·3 centimetres

TABLE IX.—NUTS.

Length of Nuts.	Breadth of Nuts. 1·7–	1·9–	2·1–	2·3–	2·5–	2·7–	2·9–	3·1–	3·3–	3·5–	3·7–	Total.
3·1–	—	—	2	—	—	1	—	—	—	—	—	3
3·3–	1	—	—	4	—	—	—	—	—	—	—	5
3·5–	1	—	—	—	1	1	—	—	—	—	—	3
3·7–	—	—	2	3	6	6	—	—	—	—	—	17
3·9–	—	—	1	4	8	7	1	—	—	—	—	21
4·1–	—	—	—	7	15	4	2	1	—	—	—	29
4·3–	—	—	—	2	16	6	7	4	—	—	—	35
4·5–	—	—	—	2	4	7	7	1	2	—	1	24
4·7–	—	—	—	—	2	8	4	1	—	—	—	15
4·9–	—	—	—	—	2	5	9	2	1	—	—	19
5·1–	—	—	—	—	—	3	3	1	1	1	—	9
5·3–	—	—	—	—	—	—	2	—	—	—	—	2
5·5–	—	—	—	—	1	—	—	—	1	—	—	2
5·7–	—	—	—	—	—	—	1	—	—	—	—	1
Total	2	—	5	22	55	48	36	10	5	1	1	158

broad, and of these 2 were from 3·1 to 3·3 centimetres long, 2 were from 3·7 to 3·9 centimetres long, and 1 was between 3·9 and 4·1 centimetres in length. Similarly, taking the bottom row but one, there were 2 nuts between 5·5 and 5·7 centimetres in length, and 1 of these was between 2·5 and 2·7 centimetres, and the other between 3·3 and 3·5 centimetres in breadth.

It will be found, on examining the table, that the long nuts are, on the whole, considerably broader than the short nuts, and if the average breadths of the nuts be calculated for various lengths, the following figures will be found:

TABLE X.

Actual Length of Nut.	Average Breadth of all the Nuts, having the Length given in the Previous Column.
Under 3·7 cm.	2·31
3·8 ,,	2·58
4·0 ,,	2·63
4·2 ,,	2·63
4·4 ,,	2·77
4·6 ,,	2·90
4·8 ,,	2·85
5·0 ,,	2·95
Over 5·1 ,,	3·04

The reader will see that, when the lengths of the nuts in a group are from 3·9 to 4·1 centimetres, we have assumed all the nuts to be exactly 4·0 centimetres long, and a similar assumption has been made in the breadths when the means were calculated. The error involved in the approximation is small, and it saves a considerable amount of trouble.

Now, it will be seen that, while the length of the nuts was increasing from about 3·1 to 5·9—*i.e.*, by about 2·6 centimetres *—the corresponding breadth was increasing by about 0·73 centimetres, which implies that an increase of 0·73 in breadth goes with an increase of 2·6 in length.

A similar method applied to the breadths gives the following results.

TABLE XI.

Actual Breadth of Nut.	Average Length of all the Nuts, having the Breadth given in the Previous Column.
Under 2·3 cm.	3·57
2·4 ,,	4·02
2·6 ,,	4·28
2·8 ,,	4·42
3·0 ,,	4·77
Over 3·1 ,,	4·78

* Not taken for quite the full distance of 2·8, because the material is very scanty at the ends. For the same reason we averaged the breadths for a few groups—*i.e.*, under 3·7 and over 5·1.

From these figures an increase in the average lengths of 1·21 centimetres is seen to correspond with an increase of, say, 1·1 centimetres in the actual breadths.

We may now examine these two results, and, remembering what has previously been said about perfect correlation, we should at first sight expect the ratio $\frac{0{\cdot}73}{2{\cdot}6}$ to measure correlation, and on the same argument we should also expect $\frac{1{\cdot}21}{1{\cdot}1}$ to do so; and it is necessary to inquire why they are unsuitable, for it is clear that the relation of length to breadth must be the same as that of breadth to length.

Perhaps the easiest way to see what has to be taken into account to correct these ratios is by asking whether we are right in taking 0·2 centimetres as the unit for both lengths and breadths. Might it not be correct to take 0·1 centimetres in one case and 0·2 in the other? This point can also be seen very easily by imagining that we are comparing, say, the number of runs made by a batsman with the rainfall on the previous day. How can we tell in such a case that a record of each single run would be the right term for comparison with each tenth of an inch of rainfall? Or if we had measured the lengths of the nuts in inches, and the breadths in centimetres, the statistical measure of correlation

ought not to be affected, but our ratios would most certainly be altered. In other words, we cannot tell that the grouping we choose arbitrarily for convenience is the one that should be used. It has been shown that, in order to make proper allowance, we must take the standard deviation as the unit of measurement; that is to say, we must measure each thing in the terms of its own standard deviation •

Now, returning to the nuts, we can apply this rule; and as the standard deviation of the lengths is 0·539, and of the breadths 0·2707 centimetre, we must remember that the unit for length is really double (almost exactly) that for breadth. Applying the method to our figures, we have—

(1*st*)	Actual increase in length - - -	$2{\cdot}6 \div 0{\cdot}54 = 4{\cdot}82$
	Corresponding increase in average breadth	$0{\cdot}73 \div 0{\cdot}27 = 2{\cdot}70$
	Ratio - - - - -	$\dfrac{2{\cdot}70}{4{\cdot}82} = 0{\cdot}56$
(2*nd*)	Actual increase in breadth - - -	$1{\cdot}1 \div 0{\cdot}27 = 4{\cdot}07$
	Corresponding increase in average length	$1{\cdot}21 \div 0{\cdot}54 = 2{\cdot}24$
	Ratio - - - - -	$\dfrac{2{\cdot}24}{4{\cdot}07} = 0{\cdot}55$

It will be seen that the two ratios are now practically equal, and give a proper measure of the correlation between the lengths and breadths of nuts. As a matter of fact, however, the method is not that which would be employed in practical work, and is mainly of use for showing the meaning

of the function used for measuring correlation, the drawback to it being that, in taking the increases in lengths and breadths, one is always more or less at a loss to know exactly what to take; and in our example we worked out the coefficient by the customary* and somewhat lengthy method, and found that its value was 0·5742, and then used the distances to give that result.

An alternative rough method would be to set out the lengths and breadths of the nuts in terms of their standard deviation in the way shown in Fig. 20, where the average breadth is shown by crosses for each length, and the space used for denoting each centimetre in length is exactly half that for the breadth, because the standard deviation is double. A line is then drawn, giving the run of the series of crosses as nearly as possible, and passing through the point where the lines representing the mean length and the mean breadth cross (see Fig. 20); it should be remembered that the crosses at the two ends, being based on few figures, need not be considered as having much weight.

* The reader who wishes to see the fuller method may refer to one of the textbooks on the subject, C. B. Davenport, 'Statistical Method' (Chapman and Hall), or W. Palin Elderton, 'Frequency Curves and Correlation' (C. and E. Layton, 1906), or to a paper by G. U. Yule in the *Journal of the Statistical Society* (vol. lx., pp. 812 *et seq.*).

In order to estimate the coefficient of correlation from such a diagram, we find the ratio of the distance of the sloping line from the horizontal line to the distance of the sloping line from the vertical line. In this case it will give the value mentioned above—namely 0·5742.

We may now take another example, and compare

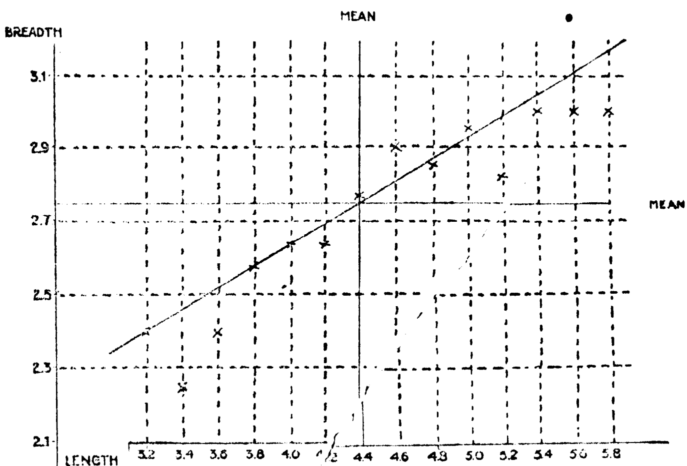

FIG. 20.—SHOWING CORRELATION BETWEEN THE LENGTH AND BREADTH OF NUTS.

the lengths and breadths of shells. The following correlation table gives the statistics, and it is clear from it that there is a very close relationship between length and breadth, which we can leave the reader to examine for himself with the help of Fig. 21. It will be seen from this figure that

Table XII.—Shells.

Breadth of Shells.	Length of Shells. 1·0–	1·3–	1·6–	1·9–	2·2–	2·5–	2·8–	3·1–	3·4–	3·7–	4·0–	4·3–	Total.
0·7–	1	1	—	—	—	—	—	—	—	—	—	—	2
0·8–	—	1	—	—	—	—	—	—	—	—	—	—	1
0·9–	—	1	2	—	—	—	—	—	—	—	—	—	3
1·0–	—	—	3	2	1	—	—	—	—	—	—	—	6
1·1–	—	—	—	6	1	1	—	—	—	—	—	—	8
1·2–	—	—	—	—	1	1	1	—	—	—	—	—	3
1·3–	—	—	—	—	1	3	2	1	—	—	—	—	7
1·4–	—	—	—	—	—	4	3	1	—	—	—	—	8
1·5–	—	—	—	—	—	—	4	7	2	—	—	—	13
1·6–	—	—	—	—	—	—	1	7	2	—	—	—	10
1·7–	—	—	—	—	—	—	—	6	8	3	1	—	18
1·8–	—	—	—	—	—	—	—	—	1	4	1	—	6
1·9–	—	—	—	—	—	—	—	—	1	3	1	—	5
2·0–	—	—	—	—	—	—	—	—	—	2	2	—	4
2·1–	—	—	—	—	—	—	—	—	—	—	2	3	5
2·2–	—	—	—	—	—	—	—	—	—	—	—	1	1
Total	1	3	5	8	4	9	11	22	14	12	7	4	100

the coefficient of correlation is 0·95, which is very near the perfect correlation expressed by unity.

It is necessary to insert a note of warning here,

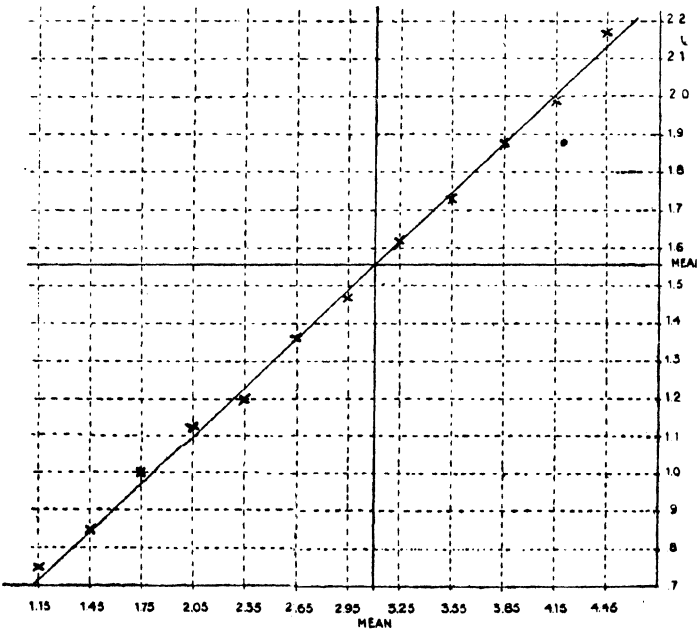

FIG. 21.—SHOWING CORRELATION BETWEEN THE LENGTH AND BREADTH OF SHELLS.

as we have up to the present assumed that we can always draw a straight line through the crosses in our diagrams; but in some cases these crosses lie on a curved line, and it is possible to get perfect correlation, even though the straight line we draw

does not lead one to expect it at first sight. To show this we will give an example:

TABLE XIII.

Size of First Thing.	Size of Second Thing.
1	0·26
2	0·52
3	1·04
4	2·07
5	4·15

Now if the reader examines Fig. 22, he will see that the method we have used so far

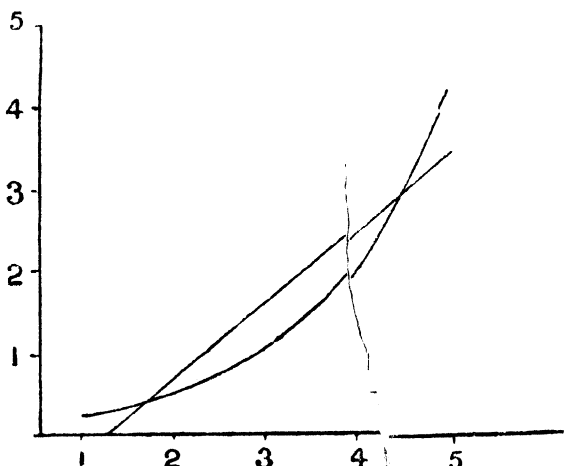

FIG. 22.

would lead to the conclusion that the measure of correlation is 0·9, whereas it should be unity,

for the series has been constructed so that the size of the second thing is absolutely dependent on that of the first. The diagram also shows the cause of the difference, for the series is not a straight line, but obviously a curved line. Here we come, therefore, to the conclusion that the method we have used hitherto is only justifiable if the crosses, when plotted out, run approximately in a straight line; if they run otherwise, the method requires modification, and the matter becomes too troublesome for an elementary book. The reader will, however, find the straight line is applicable in almost all cases, and even when it should not, strictly speaking, be used, the error is generally small.

As coefficients of correlation are used to a very great extent, it is important that anyone who wishes to understand modern work in statistics should become familiar with this coefficient, and in order to help to give an idea of the values obtained, Table XIV. may be of help.

The correlation between measurements of the human race in this table are drawn from a paper by Dr. W. R. Macdonell, 'On Criminal Anthropometry and the Identification of Criminals,'* and may be taken as an indication of the scale of

* 'Biometrika,' vol. i., p. 177, etc.

TABLE XIV.

Things Compared.	Coefficient of Correlation.
High Correlation:	
Length and breadth of shells - -	0·95
Left cubit and left middle finger -	0·85
Height and left cubit - - -	0·80
Height and left foot - - -	0·74
Medium Correlation:	
Height and left middle finger - -	0·66
Effectiveness of vaccination and resistance to smallpox - -	0·64
Length and breadth of nuts - -	0·57
Capacity of skull and greatest horizontal breadth (Naqada skulls, male) - - - - -	0·43
Height and breadth of face - -	0·35
Low Correlation:	
Length of skull and auricular height (French race) - - - -	0·29
Length of life in husband and wife	0·22
Weight of healthy heart and age of person - - - - -	0·14
Heart and spleen - - - -	0·08

relationship used. A moment's consideration will show that the order—

1. Left cubit and left middle finger;
2. Height and left cubit;

3. Height and left foot;
4. Height and left middle finger;
5. Height and breadth of face—

is the order of relationship one would expect, for as the left middle finger is part of the left cubit, one would naturally expect a close relationship between them. With regard to the second item, one has merely to remark that a tall man would seldom have a short arm, for it would look noticeably strange if he did. Again, it would seem likely that one would more often find a tall man with a short middle finger than with a short foot. In the same way breadth of face would hardly be expected to have so close a relationship to height as to the other measurements. The numerical values give precision to these rough ideas, and in order to complete the scale a few other coefficients are given. The two lowest in the scale are taken from a paper by Dr. M. Greenwood, junior,* and give examples of very low correlation.

The vaccination and smallpox result is given as a relationship which has been much discussed, but though it is not reached by quite the same method of calculation, the result is strictly comparable with the other coefficients given.

In concluding this chapter, we may point out

* 'Biometrika,' vol. iii., p. 69.

that the actual cause of the correlation cannot always be revealed by the body of statistics under examination. Taking the vaccination-smallpox problem as an example, the cause of the high correlation might be that people of the better class, living in districts where isolation was easy, were vaccinated, while those living in crowded, unhealthy districts were unvaccinated. Then, if we assume vaccination to be valueless, but admit that small-pox epidemics are more likely to occur in unhealthy districts, we might still have a high correlation between vaccination and smallpox. In this case the point has been investigated so that there seems little, if any, doubt that the real cause is the usefulness of vaccination, the example indicates the necessity of great care in interpreting statistical results. A good method may easily be spoilt by careless application or biassed interpretation.

CHAPTER VI

PROBABLE ERRORS

In Chapter I. we noticed that the means found from different samples of the same thing did not vary greatly, but we also saw that they did not always agree exactly, and this leads us to ask how far one may rely on one's calculations; because, if the means found from samples do not agree, how can we tell what would be the mean size of nuts if we took all the nuts in the world? Now the reader will remember that in Chapter II. we saw that the deviations from a mean can be measured by using quartiles, and in Chapter IV. we saw that an improved method was to use the function called the standard deviation.

If the standard deviation measures the way the sizes of nuts or shells vary, it is not very difficult to see that it should help us to decide whether we may rely on a calculated mean or not.

Now, if the standard deviation measures the way deviations occur in the size of an object, it will also measure the deviations in means—that is to

say, if we calculate a number of means and set them out in the same way as we set out the individual nuts and shells in Chapters I. and III., we can, by calculating the standard deviations, measure the deviation that may occur in a mean. The principle is exactly the same; we are merely working with means instead of nuts. As an example will help to make this clear, we made an investigation into the average number of heads that would result from tossing six coins twenty times; knowing, of course, that the most likely number is three heads. The first time we did this we found the following distribution:

Number of heads - - -	1	2	3	4	6
Number of times the number of heads occurred - - -	2	5	7	5	1

The average number of heads was therefore 2·95, and we then made another trial and found 3·05; a third trial gave 3·05 again, and we continued this operation until we had found forty-five of these means. The result is shown in Table XV.—the last two columns of which are similar to those in Table VI.

This result tells us the standard deviation of the mean in this particular experiment; but the reader will see that we cannot go on doing this kind of trial each time we calculate a mean, and we must

TABLE XV.

Mean Number of Heads from 20 Tossings of 6 Coins.	Number of Times the Mean mentioned in the Previous Column was Found.	Deviation from Mean.	Square of Deviation multiplied by Number of Times the Deviation occurred.
2·25	1	−0·75	0·56
2·55	1	−0·45	0·20
2·6	2	−0·4	0·32
2·7	2	−0·3	0·18
2·75	4	−0·25	0·25
2·8	3	−0·2	0·12
2·85	3	−0·15	0·07
2·9	3	−0·1	0·03
2·95	2	−0·05	0·00
3·0	4	0·0	0·00
3·05	4	+0·05	0·01
3·1	4	+0·1	0·04
3·2	1	+0·2	0·04
3·25	2	+0·25	0·12
3·3	4	+0·3	0·36
3·4	3	+0·4	0·48
3·55	1	+0·55	0·30
3·65	1	+0·65	0·42
Total number of means - -	45		3·50

$$\text{Average square of deviations} = \frac{3{\cdot}50}{45} = 0{\cdot}078.$$

$$\text{Standard deviation of mean} = \sqrt{0{\cdot}078} = 0{\cdot}28.$$

see if some shorter method can be found to give the same result.

Now if you turn to Chapter I., and consider any one of the diagrams, you can see that the larger the standard deviation, or, which is approximately the same thing, the larger the difference between the median and quartile, the steeper the line at the top of the figure. Now if this line rises steeply, we know that the cases vary considerably, while if the line is horizontal, there are many cases of about the same size; if, however, there are many cases like the median, and we shift the median four or five terms to the right or left, it makes only a little difference; but if only a few cases are like the median, a shift of four or five spaces may make a considerable difference in the value of the median.

Another element, however, enters into our estimate. If the mean size of a nut is calculated from 5,000 nuts, we should feel fairly certain about the result; but if we based our mean value on only five cases, we should not feel so happy about it. In other words, we must use the number of cases investigated as well as the standard deviation in estimating the accuracy of a mean which has been calculated. In order to bring in the number of cases properly, it does not follow that we must divide the standard deviation by the number of

cases investigated. It has, in fact, been shown that the accuracy depends, not on the number, but on the square root of the number of observations; so one doubles the accuracy by taking four times, and trebles the accuracy by taking nine times, the number of measurements.

Applying this rule to our example, we find from our first tossing of coins, which may be taken as the real experiment and the other tossings as confirmatory experiments to test our result, that the standard deviation of the distribution was 1·16, and dividing by the square root of the number of tossings (twenty) on which our mean depended, we get ·27 as the standard deviation, which agrees very closely with the result found from actual experiment in Table XV. In the particular case the reader can check the standard deviation from the series in the footnote to p. 35, and he should obtain 1·22 instead of 1·16, the former representing the most probable result and the latter being that which actually occurred in our experiment. We mention this point to show that standard deviations themselves and, in fact, all statistical measures are subject to deviations.

In practical work therefore, when we find a mean or other statistical function, we put after it a figure which represents the deviations we might get from the calculated value. Taking the nuts as

an example, we have found that the mean length of 185 nuts is 4·40 centimetres, and the standard deviation is 0·49 centimetres, so the deviation to which the mean itself would be liable would be 0·036 centimetres.

It has been a custom to use 0·6745 times this —*i.e.*, 0·024 (see p. 53)—and call the result the 'probable error.' On certain assumptions, the chance of a deviation being greater than the 'probable error' so found is equal to the chance of a deviation being less; consequently the calculation shows the error that is as likely as not to be exceeded. At the same time, the assumptions on which the theory rests are not always realized; they imply that the deviations follow the bell-shaped curve, which is called the 'normal curve of error,' and is, as we have already (p. 36) seen, symmetrical, tails away at each end, and has, in fact, certain very definite properties. Now we have noticed that this does not always apply, and was clearly inapplicable to the cricket scores with which we have been dealing, so we are inclined to think that it would be a good thing to give up the fraction 0·6745, and use the standard deviation divided by the square root of the number of cases. This is, however, rather a point for specialists, and the reader will be more interested to see the practical bearing of the function generally employed.

The way this probable error is used in practice is that, when the difference between two means exceeds three times the probable error, then it is considered that the difference is significant. This rule is explained most easily by a diagram.

The mean in the normal curve which is assumed in Fig. 23 is at the middle, because the curve

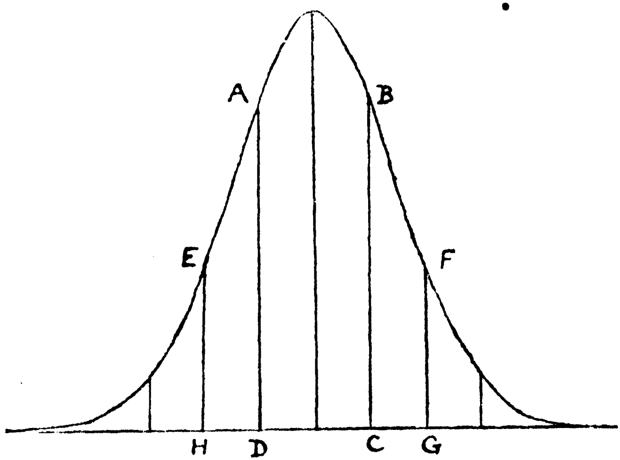

FIG. 23.

is symmetrical; the line drawn at the distance from the mean, equivalent to the 'probable error,' divides each half of the curve into half, or ABCD is half the entire area between the curve and the base. Now if we take twice the probable error, and take EGHF, we find that a large proportion of the curve is included; while if we take three times the prob-

able error, there is so little left outside that the odds against a particular result falling outside three times the probable error are 22 to 1, which, as indicated above, is very improbable.

We may now notice that all functions in statistics have their probable error, for we can never be quite sure that another selection made at random, with the object of measuring a particular thing, will give the same result as our last series of measurements, but we may feel fairly assured that it will not differ from it by three times the probable error already found.

This practical method turns up over and over again; if it is necessary to compare the sizes of skulls in two races, how else can we tell whether the differences are significant, or are merely due to the fact that we are dealing with small samples of the general population?

Perhaps the time at which one requires the 'probable error' or some equivalent function most is when we have to deal with correlation, for when the relationship between two things is not very close we may find a coefficient of, say, 0·1, which seems small; and one may wonder, if only a few cases have been dealt with, whether there is really any correlation at all, or if it is not in reality a case in which the true value is zero, but, owing to the small number of cases investigated,

the value has come out rather too large. For such cases in practice one uses the probable error, and it has been shown that the probable error of a coefficient of correlation is $0{\cdot}6745 \times \frac{1 - r^2}{\sqrt{n}}$, where r is the coefficient and n the number of cases considered;* so that if we were dealing with only 36 cases the value 0·1 would not be significant, for its probable error would be about 0·11, which is larger than the coefficient itself; while if there were as many as 100, the probable error would be 0·06, and one would still feel rather doubtful if the 0·1 meant that the two things under examination were really related.

The probable errors in the cases we have dealt with in the previous chapter are as follows:

Things Compared.	Co-efficient.	Probable Error.
Length and breadth of 185 nuts	0·57	0·03
Length and breadth of 100 shells	0·95	0·06

from which it will be clear that there is a distinct relationship in each of these cases between length and breadth.

* We should remark that when the number of cases is small this formula is not applicable; it will not give a satisfactory result if n is less than 20, and is doubtful if n is between 20 and 30.

We may here remark that the whole of this chapter depends on the assumption that the things dealt with have been taken at random, and we may say definitely that the collection of statistics in any other way is sheer waste of time. If you want to study the lengths of things, it is absurd only to take the big cases—you would never collect all the tall men in order to find the average height of a race—and yet one frequently finds people spend years of work in collecting particular cases which are absolutely unconvincing to any statistician. It is obvious in many cases, and we may take two of these to show to what wrong results people might be led. Suppose you wanted to study the size of nuts, and went to a shop which had a reputation for selling specially good nuts for the table, you would then get a very high value for both length and breadth, because the seller would have selected large specimens rather than small ones. A collection made from another shop might give a very different result; and though you would find the two lots differing in their mean value by many times the probable error found from either, you would not be justified in saying either that they were not the same sort of nuts, or that they were imported from different places.

To take a second example, we may imagine that we want to investigate the connection between the

poverty of surroundings and deformity in an individual. Now it would be absolutely useless to go into all the poor districts of London and count the number of the deformed, because there would be nothing with which to compare the result; and it would improve matters very little if we also counted all the deformed people in wealthy districts, for we might find 5,000 in the latter case and 20,000 in the former, but should have proved nothing until we had ascertained how many people there were in each district. Thus, if there were 500,000 persons residing in wealthy districts and 2,000,000 in poor districts, the two classes exhibit the same proportions. This is a simple example which we have chosen to show the need of collecting 'at random': it shows that we must not count merely the deformed, but also the number who are not deformed, if our result is to be worth anything. There are many pitfalls like these to the unwary, and no amount of knowledge of refined statistical methods can ever give value to statistics that have been collected on an unsound plan; in this connection we would add that it is far better, in working on such things, to take 5,000 or 6,000 cases 'at random,' and examine them all to see what their characteristics are, than to take 50,000 which are specially chosen; for the former tell us something, and the latter tell us absolutely nothing.

We have dwelt on this point because it is a common error, and the most serious one which anyone who tries to collect statistics can make, and we would conclude by saying that, if the reader wishes to make any investigation for himself, or even if he wishes to read any statistical work critically, he must be satisfied that the material has been collected at random from the general population of the thing investigated, whether it is nuts, or scores, or shells, or animals, or men, and has been examined without personal bias.

INDEX

THE END

BILLING AND SONS, LTD., PRINTERS, GUILDFORD